Introspection of various K-Nearest Neighbor Techniques

Mr.Suhas G. Kulkarni
Sinhgad School of Computer Studies,
Solapur, India
kulkarni.suhasg@gmail.com

Dr.M.Vinya Babu
Sinhgad School of Computer Studies,
Solapur, India
mvinayababu@gmail.com

Abstract— **K-Nearest Neighbor (kNN) technique is used in different application areas, as it is very simple, highly efficient and effective. Its main advantage is simplicity but the disadvantages can't be ignored. Hence many researchers proposed different forms of kNN technique under various situations. Broadly kNN techniques are categorized into structure based and non-structure based (structure less) techniques. The objective of this paper is to introspect the key idea, pros and cons and target data or application area behind every kNN technique. Principal Axix Tree (PAT), Orthogonal Structure Tree(OST) , Nearest Feature Line(NFL), Center Line (CL) ,k-d Tree, Ball Tree, Tunable NN etc. are structure based techniques whereas weighted kNN, Model based kNN, Ranked NN (RNN),Condensed NN, Reduced NN, Pseudo/Generalized NN, Clustered kNN (CkNN) , Mutual kNN (MkNN), constrained RkNN etc. are structure less techniques developed on the basis of kNN. Structure based techniques reduces the computational complexity and structure less methods overcome the memory limitation. Hence structure based kNN techniques can be applied to small volume of data whereas Non-structure based kNN techniques can be applied to large data set.**

Keywords*— *nearest neighbor, kNN, Data Mining.

I. Introduction

kNN is an algorithm that stores all available cases and classifies new cases based on a similarity measure. It is identified by various names like memory based reasoning, example based reasoning, instance based reasoning, case based reasoning, lazy reasoning.

kNN is used in variety of applications like classification, interpretation and predication, problem solving, function learning etc. It is also useful in different application areas like pattern recognition, text categorization, object recognition etc.

The nearest neighbor decision rule (Cover & Hart, 1967) (NN rule in the following) assigns to an unclassified sample point the classification of the nearest of a set of previously classified points. For this decision rule, no explicit knowledge of the underlying distributions of the data is needed. A strong point of the nearest neighbor rule is that, for all distributions, its probability of error is bounded above by twice the Bayes probability of error (Cover & Hart, 1967; Stone, 1977; Devroye, 1981). [1]

The variety of work carried out by different researchers in the area of kNN is been represented by Nitin Bhatia and Vandana in their survey of nearest neighbor techniques [2]. S.Dhanabal and Dr.S.Chandramathi also collected and represented some more techniques of kNN with their review of various k-nearest neighbor Query processing techniques. [3]

II. Nearest Neighbor Techniques

kNN techniques are separated in two categories :
1) Structure Based and 2) Non Structure Based (Structure Less).

A. Structure based k-NN Technique

To represent training datasets tree structures are used in structure based k-NN techniques. Ball Tree concept is been introduced by Ting Liu. It is a binary tree and constructed using top down approach in which, leaves contain information and the internal nodes are used to efficiently search through leaves. It has a better speed over kNN [4,5]. Stan Z.li and Chan K.L. introduced the concept of NFL by dividing the training data into plan.[6]

B. Non Structure Based k-NN Technique

In Structure less kNN techniques whole data is classified into training data and sample data point. Distance is evaluated from all training points to sample point and the point with lowest distance is called nearest neighbor. This technique is quite easy to implement. Bailey proposed Weighted kNN(WkNN) by adding weights to traditional kNN.[7]

III. Development in kNN

All these developments are summarized by [2] and [3]. Along with this still much of work is been carried out in this specific area of kNN.

Lailil Muflikhan and Made Putra Adnyana represented the classification of categorical data using modified k-Nearest Neighbor weighted by association rules [8]. Here accuracy is been increased by finding the relation between each attributes and class target of classification using association rules methods. And it is been used to give weight in order to calculate the distance in kNN algorithm.

As an experiment to combine the result of two or more classification methods using an evidential approach David A Bell, J.W.Gaun, and Yaxin Bi used support vector machine, kNN, and kNN Model based approach (KNNM) [9] as an extension of their own work.

Min-Ling Zhang and Zhi-Hua Zhou represented ML-kNN (Multi label) algorithm using maximum posteriori (MAP) principle to determine the label set for the new instance [10].Zacharias Vulgaris and George D Magoulas represented the different extensions of kNN for classification problems known as DB-kNN, CB-kNN, V-kNN,W-kNN etc.[11] Anders Sogaard presented a semi supervised condensed nearest neighbor for part-of-speech tagging [12]. Anupam Nath , Syed M Rahman and Akram Salah presented kNN classification (kNNC), which uses genetic algorithm (GA) [13].Yang song, Jian Huang, Ding Zhou , Hongyan Zha and C. Lee Giles proposed LI-kNN (Locally informative) and GI-kNN (Globally Informative) form of kNN[14].

Comparison of various kNN techniques is been given in the following table 1.

Sr.No.	Technique	Basic Idea	Merits	Demerits	Application Areas / Target Data
1.	k Nearest Neighbor (kNN) [1]	Uses nearest neighbor rule	1. training is very fast 2. Simple and easy to learn 3. Robust to noisy training data 4.Effective if training data is large	1. Biased by value of k 2.Computation Complexity 3.Memory limitation 4. Being a supervised learning lazy algorithm i.e. runs slowly 5. Easily fooled by irrelevant Attributes	large data samples
2	Weighted k nearest neighbor (WkNN) [2]	Assign weights to neighbors as per distance calculated	1. Overcomes limitations of kNN of assigning equal weight to k neighbors implicitly. 2. Use all training samples not just k. 3. Makes the algorithm global one	1. Computation complexity increases in calculating weights 2. Algorithm runs slow	Large sample data
3	Condensed nearest neighbor (CNN) [3,4,5]	Eliminate data sets which show similarity and do not add extra information	1. Reduce size of training data 2. Improve query time and memory requirements 3.Reduce the recognition rate	1. CNN is order dependent; it is unlikely to pick up points on boundary. 2. Computation Complexity	Data set where Memory requirement is main concern
4.	Reduced Nearest Neigh (RNN) [6]	Remove patterns which do not affect the training data set results	1. Reduce size of training data and eliminate templates 2. Improve query time and memory requirements 3.Reduce the recognition rate	1.Computational Complexity 2.Cost is high 3.Time Consuming	Large data set
5	Model based k nearest neighbor (MkNN) [7]	Model is constructed from data and classify new data using model	1. More classification accuracy 2.Value of k is selected automatically 3.High efficiency as reduce number of data points	1.Do not consider marginal data outside the region	Dynamic web mining for large repository
6	Rank nearest neighbor (kRNN) [8]	Assign ranks to training data for each category	1.Performs better when there are too much variations between features 2.Robust as based on rank	1.Multivariate kRNN depends on distribution of the data	Class distribution of Gaussian nature
7	Clustered k nearest neighbor [40]	To select the nearest neighbor from the clusters	1.Overcome defects of uneven distributions of training samples 2.Robust in nature	1.Selection of threshold parameter is difficult before running algorithm 2.Biased by value of k for clustering	Text Classification
8	Reverse k nearest neighbor [41-46]	Objects that have the query object as their nearest Neighbor, have to be found.	1. Approximate results can be obtained very fast. 2. Well suited for 2-Dimensional sets 3. Well suited for finite, stored data sets 4. Provides decision support	1. requires $O(n2)$ time 2. do not support arbitrary values of k 3. cannot deal efficiently with database updates, 4. are applicable only to 2D	Spatial data set
9	Continuous RkNN [47]	To monitor the regions upon updates using FUR tree	1. Overcomes the difficulties of using the kNN and RkNN queries on moving objects. 2. Best suited for monochromatic cases	1. Not suited for bichromatic cases 2. Not suited for large population of continuously moving objects. 3. Memory Limitation	Moving object data set

10	Constrained RkNN [48]	To find the RkNN on moving objects based on constrains	1. Communication load is minimized. 2. CRkNN can be applied to both monochromatic and bichromatic cases.	1. Approximate result can be obtained for bichromatic cases.	Moving object data set especially in GPS
11	Aggregate kNN [49]	To use aggregate function for finding the nearest neighbor	1.Provides memory-resident queries and cost models that accurately predict their performance in terms of node accesses	1. Cost for evaluating the disk-resident query model is high. 2. Lazy algorithm	Spatial data set
12	Mutual Nearest Neighbor [50]	To find the Mutual Nearest Neighbor using TB-tree.	1 .Uses batch processing and reuse technology for reducing I/O cost and CPU time. 2. HCMNN is used to reduce the searching time of all the data again and again.	1. Applied only to the monochromatic datasets. 2. Computational complexity	Moving object data set
13	Modified k nearest neighbor (MkNN) [10]	Uses weights and validity of data point to classify nearest neighbor	1.Partially overcome low accuracy of WkNN 2.Stable and robust	1.Computation Complexity	Methods facing Outlets
14	Pseudo/Generalized Nearest Neighbor (GNN) [9]	Utilizes information of n-1 neighbors also instead of only nearest neighbor	1.uses n-1 classes which consider the whole training data set	1.does not hold good for small data 2.Computational complexity	Large data set
15	Clustered k nearest neighbor [11]	Clusters are formed to select nearest neighbor	1.Overcome defect of uneven distributions of training samples 2.Robust in nature	1.Selection of threshold parameter is difficult before running algorithm 2.Biased by value of k for clustering	Text Classification
16	Ball Tree k nearest neighbor (KNS1) [21,22]	Uses ball tree structure to improve kNN speed	1.Tune well to structure of represented data 2.Deal well with high dimensional entities 3.Easy to implement	1.Costly insertion algorithms 2.As distance increases KNS1 degrades	Geometric Learning tasks like robotic, vision, speech, graphics
17	k-d tree nearest neighbor (kdNN) [23]	divide the training data exactly into half plane	1.Produce perfectly balanced tree 2.Fast and simple	1.More computation 2.Require intensive search 3.Blindly slice points into half which may miss data structure	organization of multi-dimensional points
18	Nearest feature Line Neighbor (NFL) [24]	take advantage of multiple templates per class	1.Improve classification accuracy 2.Highly effective for small size 3.utilises information ignored in nearest neighbor i.e. templates per class	1.Fail when prototype in NFL is far away from query point 2.Computations Complexity 3.To describe features points by straight line is hard task	Face Recognition Problems
19	Local Nearest Neighbor [25]	Focus on nearest neighbor prototype of query point	1.Cover limitations of NFL	1.Number of Computations	Face Recognition
20	Tunable Nearest Neighbor (TNN) [26]	A tunable metric is used	1.Effective for small data sets	1.Large number of computations	Discrimination Problems
21	Center based Nearest Neighbor (CNN) [27]	A Center Line is Calculated	1.Highly efficient for small data sets	1. Large number of computations	Pattern Recognition
22	Principal Axis Tree Nearest Neighbor (PAT) [28]	Uses PAT	1.Good performance 2.Fast Search	1.Computation Time	Pattern Recognition
23	Orthogonal Search Tree Nearest Neighbor [29]	Uses Orthogonal Trees	1.Less Computation time 2.Effective for large data sets	1.Query time is more	Pattern Recognition
24	Multi Label –kNN (ML-KNN)	Uses the set of labels for each instance in training set	1.motivated to overcome the concept of ambiguity	1. problem remains still unsolved	Text Classification
25	Density Based kNN Classifier – DB-kNN	Uses the structural density concept for evaluating the significance of each neighbor, along with the distances. -it explores the potential of evaluating neighbors rather than merely counting them.	1. performance is generally better than that of kNN.	1. much slower than kNN due to the structural density calculation	Pattern Recognition

26	Variable k Nearest Neighbor Classifier – VkNN	uses the *DC(*Degree of Certainty) concept, and estimates the optimum *k* for each classification.	1.Very Fast 2. Performs generally better than kNN	1. for very sparse datasets the optimum *k* found may not be valid and the results may not be better than those of kNN.	Pattern recognition
27	Weighted kNN Classifier – W-kNN	performs an evaluation on the features instead of on the patterns. Each feature is evaluated and assigned a weight based on how useful this feature is for discerning the classes of the dataset. To do this a new concept is introduced namely that of the Index of Discernibility (ID).	1. exceptionally fast 2.less complications 3. CPU time required is minimal for various values of *k*	1.need to calculate ID	Pattern recognition
28	Class Based kNN Classifier – CB-kNN	For every test element, the *k* nearest elements of each class are taken. value of *k* is selected to maximise the *DC and* harmonic mean of the distances of these neighbors is calculated and means are compared to select class yielding the lowest value	1. developed for datasets which are unbalanced regarding their class structure	1. large number of computations makes the overall CPU time naturally longer	Pattern recognition
29	Discernibility kNN Classifier – D-kNN	Similar to DB-kNN but considers discernibility of each element	1.proved reliable in terms of net reliability	1.not much change in time efficiency	Pattern recognition

TABLE 1

IV. Conclusion

We tried to compare various kNN techniques. A lot of work is been continuously going on in this specific area. As this method is easy to implement researchers are proposing the different forms and even making an attempt to overcome the shortfalls. The methods studied are significant in given circumstance and hold good in particular field.

References

[1] F Angiulli, "Fast Condensed Nearest Neighbor", ACM International Conference Proceedings, Vol 119, pp 25-32.

[2] N Bhatia,Vandana "Survey of Nearest Neighbor Techniques " IJCSIS Vol 8 no 2 2010

[3] S. Dhanabal, Dr.S.Chandramathi " A Review of various k-Nearest Neighbor Query Processing Techniques " International Journal of computer applications vol 31- No 7 Oct-2011.

[4] T. Liu, A. W. Moore, A. Gray, ―New Algorithms for Efficient High Dimensional Non-Parametric Classification‖, Journal of Machine Learning Research, 2006, pp 1135-1158.

[5] S. N. Omohundro, ―Five Ball Tree Construction Algorithms‖, 1989, Technical Report.

[6] S. Z Li, K. L. Chan, ―Performance Evaluation of The NFL Method in Image Classification and Retrieval‖, IEEE Trans On Pattern Analysis and Machine Intelligence, Vol 22, 2000.

[7] T. Bailey and A. K. Jain, ―A note on Distance weighted k-nearest neighbor rules‖, IEEE Trans. Systems, Man Cybernatics, Vol.8, pp 311-313, 1978.

[8] Lailil Muflikhah, Made Putra Adnyana "Classifying Categorical Data Using Modified K-Nearest Neighbor Weighted by Association Rules".

[9] David A. Bell , J.W. Guan, and Yaxin Bi " On Combining Classifier Mass Functions for Text Categorization " IEEE Transactions on Knowledge and Data Engineering, Vol 17, No.10, October 2005

[10] Min-Ling Zhang and Zhi-Hua Zhou "A *k*-Nearest Neighbor Based Algorithm for Multi-label Classification"

[11] Zacharias Voulgaris and George D. Magoulas "Extensions of the k Nearest Neighbor Methods for Classification Problems "

[12] Anders Sogaard "Semisupervised condensed nearest neighbor for part-of-speech tagging " Proceedings of the 49yh Annual Meeting of the Association for Computational Linguistics : shortpapers , pages 48-52, Portland, Oregon, June 19-24, 2011. C 2011 Association for Computational Linguistics.

[13] Anupam Kumar Nath Syed M. Rahman Akram Salah "An Enhancement of k-Nearest Neighbor Classification Using Genetic Algorithm"

[14] Yang Song, Jian Huang, Ding Zhou, Hongyuan Zha, and C. Lee Giles "IKNN:informative K-Nearest Neighbor Pattern Classification " J.N. Kok et al. (Eds) : PKDD 2007, LNAI 4702, pp.248-264,2007.c Springer-Verlag Berlin Heidelberg 2007